Estructuras de clasificación en español. Terminología y adquisición de conocimiento explícito para la Web semántica

Guadalupe Aguado de Cea
Inmaculada Álvarez de Mon
Universidad Politécnica de Madrid

Resumen

El estudio de las estructuras de clasificación que aquí presentamos responde al interés que la terminología tiene en las aplicaciones informáticas relacionadas con la Web semántica. La necesidad de dotar de contenido semántico, comprensible para el ordenador, a las páginas de Internet ha llevado a los informáticos a intentar extraer automáticamente de los textos los términos clave de un determinado campo de conocimiento y las relaciones que entre esos términos se establecen. Una etapa imprescindible para automatizar este proceso es la sistematización y localización de aquellas expresiones lingüísticas que señalan la presencia de un término e indican su relación con el resto de los términos de ese ámbito. Son muchos los trabajos que existen, tanto en inglés como en español, sobre la relación ontológica *PARTE_DE* (Charniak y Berland, 1999; Climent, 2000; Díez Orzas, 1999). Sobre la relación *SUBCLASE_DE* existen algunos trabajos en ingeniería lingüística en inglés (Hearst, 1992) y también en el ámbito de los lenguajes de especialidad (Trimble, 1985; Wignell *et al.*, 1993, entre otros). Sin embargo, los trabajos en español son casi inexistentes pese a la importancia que tienen las estructuras de clasificación.

Esta comunicación integra el estudio terminológico y la lingüística de corpus. Mediante un corpus "ad hoc" de texto didáctico y el Corpus de Referencia de la Real Academia Española (CREA), limitado al dominio de la Ciencia y la Tecnología en el español (de España) se ha procedido a sistematizar las estructuras de clasificación extraídas. En este trabajo se presentan solamente los datos relativos al lema "clasificar". La búsqueda en el corpus reveló la escasa presencia de vocablos exclusivamente indicadores de una relación de clase. La mayor parte de los verbos encontrados podían utilizarse con otros significados y sólo la presencia adicional de sustantivos, que podríamos considerar como un tipo especial de hiperónimos, como "grupo", "tipo" o "clase", además de ciertos determinantes, como "los siguientes" o los numerales indican claramente una relación *SUBCLASE_DE.*

Palabras clave: Estructuras de clasificación, adquisición de conocimiento, terminología.

Introducción

La necesidad de dotar de contenido semántico, comprensible para el ordenador, a las páginas de Internet ha llevado a los informáticos a desarrollar ontologías de forma automática. Muchas de las técnicas empleadas en este proceso utilizan lenguaje natural. Una etapa imprescindible es la sistematización y localización de aquellas expresiones lingüísticas que señalan la presencia de un término e indican su relación con el resto de los términos de ese ámbito. El estudio de las estructuras de clasificación que aquí presentamos contribuye, por un lado, a establecer relaciones entre conceptos y, por otro, responde al interés que la terminología tiene en las aplicaciones informáticas relacionadas con la Web semántica. Como metodología de trabajo integra los estudios de terminología y la lingüística de corpus.

La *web* semántica y las ontologías

La revolución que ha supuesto Internet como medio de acceso a la información, o al conocimiento, ha sido enorme. No obstante, los ingenieros e informáticos

continuamente imaginan nuevas aplicaciones y posibilidades. En este sentido, Tim Berners Lee, el científico que ideó la red mundial promueve en la actualidad un nuevo avance, que él denomina la *Web semántica* (Berners-Lee y Fischetti, 1999). La *Web* semántica pretende que los documentos de Internet estén anotados con información sobre su contenido de modo que pueda ser reconocida por el ordenador. El lenguaje que utiliza Internet, el HTML, sólo da información sobre los aspectos visuales o formales de los documentos, como el tipo de letra o las imágenes que incluye, pero no aporta ningún dato sobre el contenido informativo del texto.

Para hacer viable la web semántica han surgido las ontologías que organizan el conocimiento de un dominio de forma comprensible para un ordenador. El término "ontología" se toma de la filosofía. En esta disciplina, se utiliza para hacer referencia a la esencia misma del ser, a su existencia, de *onto,* ser. En Inteligencia Artificial, se ha utilizado porque, para los sistemas de Inteligencia Artificial, lo que "existe" es lo que puede representarse de forma computacional. Se trata, en definitiva, de una descripción formal de los conceptos pertenecientes a un dominio y de las relaciones entre los conceptos. Este conocimiento tiene que estar consensuado y ser reutilizable. Según (Gruber, 1993), en una de las definiciones más citadas, "una ontología es una especificación explícita de una conceptualización" y, en palabras de (Studer et al., 1998): "*… a formal, explicit specification of a shared conceptualization".* Incluye, por consiguiente, un vocabulario de términos y la especificación de su sentido.

Para representar el conocimiento de un dominio concreto, una ontología consta de conceptos o clases, relaciones, funciones, instancias y axiomas. Los conceptos o clases son las nociones que se intenta formalizar. Los conceptos pueden ser clases de objetos, métodos, planes, estrategias, procesos de razonamiento, etc., dependiendo del tipo de ontología. En una ontología de cine, por ejemplo, se distinguen como conceptos, la película, el director o directora, el guionista, el autor del libro en el que se basa la película, el productor, los actores o los personajes, entre otros. Las relaciones representan la interacción y el enlace entre los conceptos del dominio, y de esta forma crean la taxonomía del dominio. Por ejemplo: *subclase-de, parte-de, parte-exhaustiva-de, conectado-a,* etc. Las funciones son un tipo concreto de relación en la que se identifica un elemento mediante el cálculo de una función que afecta a varios elementos de la ontología. Por ejemplo, pueden aparecer funciones como *categorizar-clase, asignar-fecha*, etc. Las instancias representan las diferentes realizaciones de una clase o concepto. Así, en la ontología de cine, es necesario distinguir entre las instancias correspondientes a los actores, a los directores, realizadores, etc. Un ejemplo de clasificación de instancias mediante patrones lingüísticos es el trabajo de Cimiano y Staab (2004), quienes abordaron la clasificación de entidades o clases geográficas mediante búsquedas en Google empleando los patrones lingüísticos de Hearst (1992). Por ejemplo, Amsterdam es una instancia de la clase ciudad y Atlántico una instancia de la clase océano. Los axiomas son teoremas que se declaran sobre relaciones que deben cumplir los elementos de la ontología. Por ejemplo: "*Si A y B son subclase de C, entonces A no es subclase de B*". Los axiomas permiten, junto con la herencia de conceptos, inferir conocimiento que no esté indicado explícitamente en la taxonomía de conceptos. Cuanto mayor número de axiomas incluye una ontología, su formalización es mayor.

Hoy en día, existen muchas ontologías y de tipos muy variados en función de la aplicación a la que se destinen. Un concepto puede representarse de muchas formas y,

por consiguiente, pueden coexistir múltiples representaciones de un mismo concepto. Las definiciones dependen del lenguaje empleado en su descripción y también de los fines con los que se realiza la ontología. Por ello, el objetivo es establecer relaciones entre los elementos de una o más ontologías, para descubrir conexiones, especializaciones, generalizaciones, etc. Esto se denomina *mapping*. Se trata de contrastar entre sí las ontologías existentes para establecer sus coincidencias, sus diferencias y acceder a aquélla que interesa más. Este procedimiento de comparación se realiza de diversas formas, pero, en la actualidad, se ha visto la conveniencia de recurrir a técnicas que incorporan lenguaje natural (Van Hager *et al.*, 2005).

Adquisición de conocimiento y lingüística de corpus

El objetivo de nuestro trabajo se enmarca en el ámbito general de la adquisición de conocimiento, aunque también está relacionado en parte con la extracción de información, y tiene una doble finalidad: establecer los patrones de clasificación del español para, después elaborar reglas que permitan la automatización del proceso. En el ámbito de la extracción de información y la adquisición de conocimiento, es frecuente utilizar las páginas de Internet (Gelbuck *et al.*, 2002; Smarr y Grow, 2002; Cimiano y Staab, 2004) como corpus virtual. Sin embargo, en este caso, no se ha utilizado a Google, como defienden algunos (Volk, 2002), ya que se reveló como una herramienta poco adecuada. Una prueba con el verbo "clasificar", que podríamos considerar fundamental, devolvió tan sólo ejemplos de su significado en el ámbito del deporte. Se consideró, por ello, importante partir de un corpus *ad hoc*, que se definiría, siguiendo a (Pearson, J., 1998), como un *special purpose corpus, a corpus whose composition is determined by the precise purpose for which it is to be used.* De esta forma, se evita uno de los principales inconvenientes de los *corpus* a la hora de extraer conocimiento lingüístico y es la escasa frecuencia de aparición de los patrones o construcciones de interés.

Para elaborar dicho corpus *ad hoc*, nos basamos en las siguientes ideas en torno a la clasificación. En primer lugar, la clasificación es una función propia de lo que podríamos definir como los géneros prototípicos de acceso al conocimiento científico-técnico, los libros de texto o manuales y las enciclopedias. Los estudios existentes sobre este tipo de discurso como el de Trimble (1985) y Wignell *et al* (1992) confirman que en el texto expositivo de carácter didáctico, la función de la clasificación es esencial. En segundo lugar, en algunos dominios, la clasificación constituye la actividad fundamental. Determinadas disciplinas científicas tienen como objetivo primordial el establecimiento de taxonomías del conocimiento: la zoología clasifica los animales y la fitología, las algas, etc. Por ello, para nuestro corpus *ad hoc* se partió de un manual clásico de microbiología y parasitología (Matilla *et al.* 1978), que unía a su condición de libro de texto, el pertenecer a una disciplina en la que la clasificación es fundamental. A partir de las construcciones extraídas de este corpus *ad hoc* se estableció la lista de palabras clave o palabras "semilla". Para verificar e incrementar los patrones lingüísticos se utilizó el corpus de la Real Academia, el CREA, Corpus de Referencia del español actual, pero limitado al dominio de la Ciencia y la Tecnología en el español (de España). De esta forma, se evitaba la posible ambigüedad que podía derivarse del uso de las variantes diatópicas, diafásicas y diastráticas. Los ejemplos de uso del CREA se utilizaron para establecer los patrones sintácticos.

La relación ontológica *SUBCLASE_DE*

En este trabajo nos centramos en la relación *SUBCLASE_DE*. Se trata de una relación esencial en el ámbito de las ontologías, ya que para toda entidad, concepto o clase de una ontología podría encontrarse un término "padre". Además, resulta clave para el concepto de "herencia", por el que una subclase hereda las propiedades que estaban definidas para la superclase. En los estudios semánticos En los estudios semánticos recibe el nombre de hiponimia (Eagles, 1999) porque el significado de una palabra está incluido en el significado de otra, de carácter más amplio. (Cruse, 1986) la denomina "taxonómica". Desde el punto de vista de la lógica, se trata de una relación de inclusión. En el ámbito de la computación, los patrones lingüísticos de clasificación más utilizados en la extracción de información a partir de texto se establecieron en el trabajo clásico de (Hearst, 1992). Este tema se había abordado ya en el ámbito del IFE y los lenguajes de especialidad, desde una perspectiva didáctica (The British Council, 1979; Trimble, 1985; Wignell *et al.,* 1993).

Resultados

Los patrones hallados se han sistematizado en función de las diferentes variantes (*word-forms*) del lema principal. En este trabajo, nos vamos a centrar en el lema "clasificar" y en sus formas verbales que indican la existencia de una relación de *SUBCLASE_DE*:

1. clasifica
2. se clasifica
3. se clasifican

Para cada una de las variantes, se estableció su régimen de uso sintáctico-semántico.

1. CLASIFICA

1.1 AGENTE **clasifica H en** N (tipos/grupos/clases) **según/ de acuerdo con** (criterio)
AGENTE **clasifica** H **según/ de acuerdo con**(criterio)

2. SE CLASIFICA

2.1 H se clasifica **como** X o Y
2.2 **X se clasifica dentro de** la familia o del grupo H, **con la denominación**...

3. SE CLASIFICAN

3.1 **Según/ de acuerdo con** (criterio), las/los H **se clasifican en** X, Y, Z, etc.
3.2 **se clasifican en** H de X o de Y
3. 3 **Los/las H se clasifican en:/ en los siguientes grupos:**
3.4 ..., **se clasifican** (generalmente / básicamente/ comúnmente) **en** diversos/varios tipos.
3.5 Los/las H **se clasifican en** X o Y
3.6 **Se clasifican** los/las H **en** N/grandes/diversos/ grupos...
3.7 Los/las H **se clasifican como** X/ X e Y

A partir de los ejemplos extraídos del CREA para cada una de estas expresiones, y de los esquemas de patrones aquí expuestos, se enunciaron una serie de heurísticas. Estas heurísticas se elaboraron siguiendo el patrón "si..., entonces ... ". Pueden, por tanto, constituirse en reglas, ya que primero se presenta la constante, el antecedente, y luego la variable, el consecuente. Mediante la formalización computacional de estas reglas, se podría lograr el aprendizaje automático de los hipónimos correspondientes a un hiperónimo o del hiperónimo propio de una serie de co-hipónimos. La presencia en un texto de una de las realizaciones del lema "clasificar", pone en funcionamiento una condición que deberá formalizarse computacionalmente.

Si CLASIFICA/ SE CLASIFICA / SE CLASIFICAN aparece en un texto del dominio del DEPORTE, subgénero NOTICIAS DEPORTIVAS, entonces no indica la relación *SUBCLASE_DE.*

Si CLASIFICA/ SE CLASIFICA / SE CLASIFICAN van seguidos en un contexto próximo de GRUPOS/TIPOS/CATEGORÍAS u otros sinónimos, entonces los verbos señalan una relación *SUBCLASE_DE.*

Si CLASIFICA/ SE CLASIFICA / SE CLASIFICAN detrás de la preposición EN aparece un número CARDINAL, entonces éste indica la cantidad de tipos o hipónimos que se identifican.

Si a CLASIFICA/ SE CLASIFICA / SE CLASIFICAN siguen (dentro de la misma oración) las preposiciones EN o COMO seguidas de una cadena de unidades (letras o palabras) separadas por Y, entonces hay enumeración de 2 hipónimos.

Si a CLASIFICA/ SE CLASIFICA / SE CLASIFICAN siguen (dentro de la misma oración) las preposiciones EN o COMO seguidas de una cadena de 2 ó más unidades (letras o palabras) separadas por una coma "," entonces hay enumeración de más de 2 hipónimos. El último hipónimo de la cadena aparece precedido de E/ Y/ O ó seguido de ETC,.

Conclusiones

La búsqueda en el corpus de las palabras clave para señalar la presencia de una relación *SUBCLASE_DE* entre dos términos reveló que la relación de clase se descubre por la presencia adicional de algunos sustantivos genéricos, que podríamos considerar como un tipo especial de hiperónimos: "categoría", "clase", "familia", "grupo", "orden", "rama", "tipo". Asimismo resultaron claves ciertos determinantes, como "los siguientes" o los numerales. Finalmente, la recuperación de hipónimos e hiperónimos se ve facilitada por la presencia de determinadas marcas tipográficas.

: indica el comienzo de la enumeración de las subclases.
, indica la presencia de otra subclase
y/ así como indican la presencia de la última subclase enumerada en esa oración
e se utiliza cuando la denominación empieza por i.
,etc. indica que la enumeración no es exhaustiva

Agradecimientos

Esta investigación está financiada, en parte, por el Ministerio de Ciencia y Tecnología (MCyT) a través del proyecto SEMANTIC SERVICES: "Infraestructura tecnológica de servicios semánticos para la *web* semántica", Plan Nacional de I + D + I 2004-2007, 2660 y, en parte, por el proyecto europeo del Sexto Programa Marco, "NeOn: Lifecycle support for networked ontologies", FP 6-027595.

Referencias

Berners-Lee, T. y Fischetti, M. (1999). "Waving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor", W3C, http://www.w3.org

Charniak, E., y Berland, M. (1999). "Finding parts in very large corpora". Proceedings of the 37th Annual Meeting of the Association for Computacional Linguistics (ACL), 54-64.

Cimiano, P. y Staab, S. (2004). "Learning by Googling". SIGKDD Explorations Newsletter 6, 2: 24-33.

Climent Roca, S. (2000). "Individuación e información Parte-Todo. Representación para el procesamiento computacional del lenguaje". *Estudios de Lingüística Española.* Volumen 8. http://elies.rediris.es/.

Cruse, D.A. (1986). "Lexical Semantics". Cambridge: Cambridge University Press.

Díez Orzas, P.L. (1999). *La relación de meronimia en los sustantivos del léxico español: Contribución a la semántica computacional.* Tesis Doctoral. http://elies.rediris.es/

Gelbuck , A., Sidorov, G. y Chanona, L. (2002). "Corpus virtual, virtual: Un diccionario grande de contextos de palabras españolas compilado a través de Internet". IBERAMIA. Workshop on Multilingual Information Access and Natural Language Processing. 7- 13.

Gil, Y., Motta, E., Benjamins, R., y Mussen, M. (eds) (2005) The Semantic Web. Fourth International Semantic Web Conference (ISWC 2005) LNCS, Springer-Verlag.

Gruber, R. (1993) "A translation approach to portable ontology specification". *Knowledge Acquisition* 5, 199-220.

EAGLES 1999 EAGLES LE3-4244: Preliminary Recommendations on Semantic Encoding. Final Report http://www.ilc.pi.cnr.it/EAGLES/EAGLESLE.PDF

Halliday, M.A.K. and Martin, J.R. (1993). W*riting Science Literacy and Discursive Power.* London: The Falmer Press.

Hearst, M.A. (1992). "Automatic acquisition of hyponyms from large text corpora". Proceedings of the 14th conference on Computational Linguistics, USA: New Jersey, Morristown, 539-545.

Pearson, J. (1998). *Terms in Context.* Amsterdam: John Benjamins Publishing Co.

Matilla, V. *et al.* (1978). "Microbiología y parasitología". Madrid: Amaro. 5ª ed.

Real Academia Española: Banco de datos (CREA) [online]. Corpus de referencia del español actual. http://www.rae.es.

Smarr, J. y Grow, T. (2002). "GoogleLing:The Web as a Linguistic Corpus". Disponible en http://www.stanford.edu/class/cs276a/projects/reports/jsmarr-grow.pdf

Studer, R., Benjamins, R., y Fensel, D. (1998). "Knowledge Engineering: Principles and Methods". Data and Knowledge Engineering, (DKE) Vol. 25, 1-2: 161-197.

The British Council (1979). *Reading and Thinking in English. Discovering Discourse*. Oxford: Oxford University Press.

Trimble, L. (1985). *English for Science and Technology*. Cambridge: Cambridge University Press.

Van Hage, W. R., Katrenko, S., y Schreiber, G. (2005). "A Method to combine linguistic Ontology-Mapping Techniques" en Y. Gil *et al.* (eds) ISWC 2005, LNCS 3729, 732-744.

Volk, M. (2002). "Using the Web as a Corpus for Linguistic Research" en Paujasalu, R. y Hennoste, T. (eds) Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Halsdur Oim. Publications of the Department of General Linguistics 3. University of Tartu. http://www.ifi.unizh.ch/cl/volk/papers/Oim_Festschrift_2002.pdf

Wignell, P., J.R. Martin and S. Eggins. (1993) "The discourse of Geography: Ordering and explaining the experiential world" in Halliday, M.A.K y Martin, J.R, 136-165.